Education Inequality during the COVID-19 period

Predicting Individual Study time for K-12 Students with Machine Learning Models

by

Jiayi Gao

An honor thesis for the Honors in the Major Statistics

at the
University of Wisconsin-Madison
2022

Abstract

Education Inequity was exacerbated at the beginning of the COVID-19 pandemic. This study aims to explore and interpret the features that contribute to education inequity in the pandemic period. The processed dataset in this study contains188,768 observations and 21 features. The author applies four widely used machine learning methods to make predictions on the weekly independent study time. There are four different settings of the dataset that was created by classifying the processed dataset with different criterion. The author further evaluates and interprets the results of these models in order to make comparisons between different settings. The results indicate that no matter how to classify the dataset, household income is always the feature that influences the independent study time the most. Overall, the research will help the parents and educators to find features that need extra attention during the pandemic period to avoid negative study performance in K-12 students.

1 Introduction

The outbreak of COVID-19 at the beginning of 2020 has brought countless trouble to people globally, including in the field of education. To avoid further spread of the virus, the U.S government set quarantine policies to restrict the social activities of the citizens, including going to school. Implementing this policy, schools shut down physically and changed to online teaching at the beginning of the COVID-19 period. With the sudden change of teaching form from offline to virtual, teachers had to teach online before they became familiar with the teaching form. What is worse is that they were having hard time designing suitable pedagogy due to the inequity in educational resources. Consequently, students were not able to get enough guidance on virtual learning, which leads to less motivation in studying [4]. Though schools and teachers strive to provide help, it was unavoidable for some students to experience learning loss. In addition, students showed a decreasing trend in study performance, especially on readings, due to the learning loss during the physical disruption from school [13]. In this case, it is important to explore more deeply what factors could contribute to possible learning loss during the COVID pandemic challenge, as it could reveal hidden education inequity.

The paper will present various classification algorithms and I will make comparisons of the results of these algorithms in the discussion section. However, the accuracy of the models is not the ultimate goal for this project, as it is complicated to make prediction on human behavior, especially under this stressful pandemic period. Instead, the project aims to explore and interpret the differences between different settings and models. More importantly, I plan to examine possible factors that negatively contribute to the individual study time of the student. The data applied in this research is from United States Census Bureau. I first calculate the features importance and extract the 10 out of 22 features that would best support our model training. Then, I put the processed data into four widely used machine learning models, which are Logistics regression, decision tree, random forest, and naive Bayesian. I also use confusion matrix to evaluate these models. The analysis will introduce people to education inequity during the COVID-19 period and show parents and educators factors that they should pay attention to in order to avoid negative education outcomes in K-12 students.

2 Literature Review

It has long been established that student family socioeconomic status is positively correlated with various educational resources and outcomes. Lucas [11] mentioned that socioeconomic advantaged parents spend more time and resources on the education of their children including keeping tracking with schools and teaching children with their own experiences. For example, in Raftery and Hout's paper [3], the prestige of father's occupation positively influences education achievement of the children. In this case, students who are from socioeconomic disadvantaged family may experience inequity in education resources. In addition, the COVID pandemic exacerbates the education inequity that influenced by parents' socioeconomic status, which was considered the biggest factor that contribute to education inequity between different races before the pandemic period [6, 8].

There is also evidence shown that with less time spend on studying, students present a worse study performance [7]. Considering factors that may influence the studying time, the education inequity might increase during COVID pandemic period. For example, as mentioned by Aguirre et al in their work [10], schools are considering reopening; students were under great stress due to the changes of class forms; teachers were burnout due to the increasing workload; parents are facing big challenges of possible unemployment. These are all possible factors that could negatively influence the studying time of students.

Scholars have also revealed evidence of education inequity during the COVID-19. More specifically, Frohn[10] finds that living condition is one of the major factors that could indirectly influence students' individual learning time, as some students who are not able to study on their own mentioned that the limited space in the house and noisy backgrounds are big obstacles for them to concentrate on learning. In addition, previous research has examined education inequity under another pandemic, the Ebola pandemic from 2013 to 2016. For example, William (2021) discussed the reflection of the Ebola pandemic and suggests that governments and schools should take special care to secondary-aged youth from poor families because they are the ones that most likely to drop out from school during the pandemic period. This provides us insights on education planning for the post pandemic period.

3 Method

This study uses four machine learning models to examine disparities in student learning time during the pandemic. Here below are the details regarding the models.

3.1 Logistic Regression

To find the result of probability of a specific event, logistic regression could be used, as it is a simple and widely used statistical method. The predictor of the specific event should be a binary variable with two opposite response, for example, yes/no and true/false. Generally speaking, the logistic regression method builds models by performing a logistic function, as shown below.

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdot + \beta_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdot + \beta_{ip})}$$

Where $\pi_i = x_{i1}, x_{i2}, ...x_{ip}$. Since Maximum Likelihood estimation reduces the error in the predicted probability, the logistic regression algorithm uses MLE to find the best coefficient from the training data. MLE is applied in the algorithm to avoid situations when the best coefficient are too close to 0 or 1. However, since the logistic regression model assumes there are no noises in the dataset, the dataset should be carefully cleaned and filtered before applying the model.

3.2 Decision Tree

Since most of the features in this research are in categorical forms, decision tree, a classifier, will be applied. Decision tree is an effective and easy-interpret algorithm. The model aims to create a training model that uses given features to make predictions on class label by exploring possible decision rule in the information provided in the dataset. Starting from the root of the tree, which is the entire data set, the decision algorithm split the information into two or more nodes and continued splitting the rest information from each node, until the sample from each node belong to the same class [12]. However, sometimes for large data set, the algorithm may provide a tree with too many nodes, which might cause overfitting. To prevent this, we should set a criteria for the maximal depth of the tree. In this case, Gini impurity will be used in order to avoid the overfitting of the model, and the function is shown below.

$$\text{GAIN}(D, x_j) = H(D) - \sum_{v \in values_{x_j}} \frac{|D_v|}{D} H(D_v)$$

where D is the training set at the parent node, and D_v is a dataset at a child node upon splitting.

3.3 Random Forest

Random Forest model is the combination of the outputs of a series of independent decision tree models. The result shown in the random forest model is the output of class that selected by most trees. Random Forest algorithm will randomly generate a subset of features that make sure the correlation between each tree stays low so that they are not relying on each other. On the other hand, decision tree algorithm will take every feature splits into consideration.

3.4 Naive Bayesian

Bayesian theorem is widely applied in the calculation of probability. The Naive Bayesian method is a supervised classification algorithm that assume all the predictors are independent. Here below is the basic formula for Naive Bayesian classifier:

$$P(c|x) = \frac{P(x)P(c)}{P(x)}$$

where $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes); $P(x)$is the prior probability of class; $P(x|c)$ is the likelihood which is the probability of predictor given class; and $P(x)$ is the prior probability of predictor.

4 Analysis

4.1 Data Source

The primary data set that will be used in this project is the "Household Pulse Survey" [1] from the United States Census Bureau. All the features contained in this data set are collected from the survey that was designed to access how U.S citizens' lives were influenced by COVID-19. The data sets were updated weekly from April 23rd, 2020, to July 21st, 2020, and biweekly until the present. The respondents that have been surveyed are randomly selected from current US citizens. Another data set that contains the unemployment insurance policy during the COVID period provided by CUSP (COVID-19 Us State Policy) [2] will also be used in this research. We merged this data set with the House Pulse Survey data set by state.

To match the research interest for education inequality in the most severe pandemic period, I extracted data from week 14 to week 30, which covers information regarding the two semesters in the time period from September 2nd, 2020, to May 24th, 2021. Until this step in the research, there are 1,346,249 pieces of record and 315 features in the raw data set. I further narrowed the data set to the size of 249,110 observations by removing households that do not have K-12 students in their homes, as the research focuses on the education inequality of K-12 students. Features that might have both positive and negative influences on student study time were selected to reduce the dimension of the data set. In my feature selection process, some of the chosen features such as changes of teaching forms; time spend on live contact with teach; level of anxiety and worry of parents; students' individual study time were based on Aguirrer's paper[10]. See appendix for detail explanation for each variable in the data.

4.2 Limitation

There are some limitations existing in the data set. In the traditional discussions of education inequity, race and gender are two factors that are frequently referred to.

However, in this project, since each observation uses individual households as a unit in this data set, it is not likely to take the race and gender of the K-12 students into consideration. Another thing to be noticed is that the project uses only individual study time as the estimator to reflect the education inequity among K-12 students, and it is not likely for kindergarten students to study independently. In this case, 20% of individual study time is equal to 0, which means there are around 20% of students who do not study independently at all. However, among those who did not spend time on studying independently, there are also high school students who do not want to or do not have the resources to support themselves to study. So, it is not reasonable to remove those observations with 0 values out of the data set. Further research focused on high school students will be done in the future.

4.3 Data Processing

In this research, I used R to process and clean the dataset. To take the United States COVID Policy into consideration, I merged the data from CUSP into the original data set by the feature “states”, as the policy varies between each state. There are also values of -88, -99, and null in the data set, which stand for “question did not see”, “question seen but did not respond”, and no values respectively. I first removed all the observations with -88 and null values in the dataset. However, variables regarding teaching forms, which are teach_cancelled, teach_distance_online, teach_distance_paper, teach_distance_no.change, need extra attentions. In particular, observations that have -99 for all of the four teaching form variables were deleted because the respondent did not tell us anything about the teach form of the student(s) in the household. Then I further cleaned the teaching form variables in the data set by converting the -99 into 0, which makes these columns binary variables. The reason I change -99 into 0 is that whether applying a certain teaching form is a “yes” or “no” question and should be labeled as 0 and 1. For other variables, the -99 are considered null values and the observations with null value are removed from the data set. At this time, there are 205,568 observations

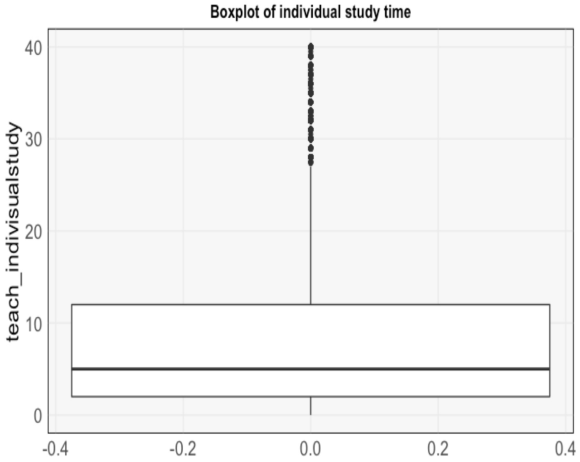

Figure 1: Box plot of independent study time of students

left in the data set. Also, by generating a box plot(Figure 1 below) for the processed data, we can see that the observations of individual study time above 28 hours per week would be outliers, which will be removed from the data set. Finally, 85% of the observations was removed from the raw dataset, and 188,768 observations were left to support our model training.

4.4 Data summary

To have a general idea of this large data set, I have done basic summaries of the individual study time of each student per week. To take an overall look, most students spend less than 10 hours per week studying on their own. The median of the data is 5, and the mean of the data is 6.99. This means the data is left-skewed, which makes sense because there are 28,417 observations (15%), who do not study on their own at all. A possible explanation would be that the motivation and ability for study of kindergarten and elementary students might not be as high as the high school students. Based on the differences between study ability and motivation, it is reasonable that the standard deviation of the entire data set is 6.51, which is relatively high. More details regard data summary were present in table 1.

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max
UI_quarantined	0.00	1.00	1.00	0.85	1.00	1.00
UI_highrisk	0.00	0.00	0.00	0.29	1.00	1.00
Marital_Statusk	1.00	1.00	1.00	1.762	3.00	5.00
Food_Sufficency	1.00	1.00	1.00	1.45	2.00	4.00
Private_Health_insurance_status	1.00	1.00	1.00	1.18	1.00	4.00
Public_Health_insurance_status	2.00	2.00	2.00	1.88	2.00	3.00
anxiety	1.00	1.00	1.00	2.26	3.00	4.00
worry	1.00	1.00	2.00	1.975	3.00	4.00
depression	1.00	1.00	2.00	1.82	2.00	4.00
no_interest	1.00	1.00	2.00	1.84	2.00	4.00
Incomeloss	1.00	1.00	2.00	1.55	2.00	2.00
Householdincome	1.00	4.00	5.00	5.08	7.00	8.00
teach_cancelled	0.00	0.00	0.00	0.223	0.00	1.00
teach_distance_online	0.00	0.00	1.00	0.70	1.00	1.00
teach_distance_paper	0.00	0.00	0.00	0.16	0.00	1.00
teach_others	0.00	0.00	0.00	0.15	0.00	1.00
teach_no.change	0.00	0.00	0.00	0.11	0.00	1.00
teach_livecontact	1.00	3.00	4.00	3.432	4.00	4.00
teach_indivisualstudy	0.00	2.00	5.00	9.30	12.00	40.00
teach_studytime	1.00	1.00	2.00	2.45	3.00	5.00
Health_insurance	0.00	1.00	1.00	0.93	1.00	1.00

Table 1: Data summary

4.5 Feature Selection

Even though we included all the features that we think may have great influences on the individual study time of each student based on our literature review and common sense, there is no way to guarantee all of these 24 features will make equal contributions to our model building. Thus, to exclude potential noise(s) in this data set, it is necessary to calculate the feature importance towards the output variable. In this project, I applied

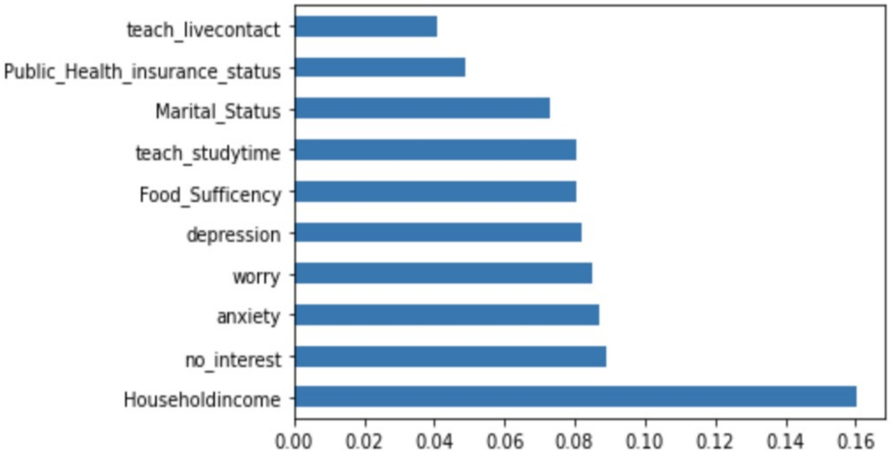

Figure 2: Feature Importance

the feature importance function in decision tree model to find the top 10 features that have greatest influences on our estimator. These features will be selected for further use in the model training process. As shown in the feature importance graph below (Figure 2), the selected 10 features are public health insurance status; levels of hours for virtual contact with the teacher; the marital status of parents in the household; food sufficiency level; levels of mental health for each parent including whether they feel anxiety, depression, worry, or no interest; level of student's study time each week compared to the time before the COVID pandemic; as well as household income, which seems to have the greatest influence on the model.

4.6 Model Training

Before I move on to the model training process, I split the data into a testing set and a training set, with 30% and 70% of the entire data set respectively. The function that I used in this project to split the data is the train test split method in learning. Since the accuracies of the prediction vary largely when I change the way I split the data, I decided to present all results for different methods and make a comparison between these results as well as between these methods. I classified the dataset under four different standards. In "case 1", I want to make a prediction on whether the students' study individually or not. So, I labeled the data by 0 and 1, which represents individual study time equal to 0 and individual study time is not 0 respectively. In “case 2”, I tried to make prediction on whether the individual study time of the student is above average or below the average. In this case, the data was labeled with 0 and 1, which represents individual study time below or equal to 7 hours per week and above 7 hours per week respectively. I classify the data by its quantile in “case 3”, so that each label could have even amount of data to support the analysis. To further explore whether the selected features have influence on higher level individual study time, I extract the part of the data above the average and labeled them by its quantile. In “case 4”, the dataset only contains 76,773 observations. Since “case 3” and “case 4” have multiple level labels, logistic regression is not applied in these two cases.

4.6.1 Logistic Regression

To achieve logistic regression, I import the ”LogisticRegression” function from the scikit-learn library.No hyperparameters were changed. The penalty was left as the default method of L2 regularization, the solver type was left as ‘lbfgs’, and C, the inverse of regularization strength, was left at 1. However, the cases with multiple level labels cannot apply logistic regression, this section only contains test results for “case 1”and “case 2”. In “case 1”, the test accuracy for the logistic regression model is 84.8% with MSE of 0.15. The test accuracy for “case 2” is 63.22% with MSE of 0.37. The detailed test results are recorded in table2 below.

Model	Accuracy	MSE
case 1	84.8%	0.15
case 2	63.22%	0.37

Table 2: Test results for logistic regression

4.6.2 Decision Tree

I applied the ”DecisionTreeClassifier” function in the sklearn. tree package to run the model. In “case 1”, the final accuracy for the test set 84.79% with the best parameter

case	Accuracy	Hyperparameter	precision	recall	F1
case 1	96.97%	max_depth = 5	pos = 0.30 neg = 0.85	neg = 0.00 neg = 1.00	0.46
case 2	63.76%	max_depth = 6	pos = 0.53 neg = 0.65	pos = 0.12 neg = 0.74	0.48
case 3	32.43%	max_depth = 8	1 = 0.36, 2 = 0.26 3 = 0.28, 4 = 0.33	1 = 0.63, 2 =0.24 3 = 0.16, 4 = 0.28	0.29
case 4	46.83%	max_depth = 5	1 = 0.48, 2 = 0.00 3 = 0.39, 4 = 0.00	1 = 0.91, 2 =0.00 3 = 0.11, 4 = 0.00	0.20

Table 3: Test results for Decision Tree

of 'criterion' = 'entropy', 'max depth' = 5. To further explore the information of the predictions made by the decision tree model, I created a confusion matrix by applying the "confusion_matrix" function. The precision is 0.30 and 0.85 for positive and negative tweets respectively and the recalls are 0.00 and 1.00. I have the macro average F1 score is 0.46. The test accuracy for "case 2" is slightly lower, which is 63.76% with macro average F1 score of 0.48. The precision shown in the confusion matrix is 0.65 and 0.53 for 0 and 1, and the recalls are 0.94 and 0.12 for 0 and 1. Compared to the previous two cases, "case 3" and "case 4" have relatively low test accuracies of 32.43% and 46.83% respectively with macro average F1 scores are 0.29 and 0.2 for "case3" and "case4". The precision and recalls are listed in above table 3.

4.6.3 Random Forest

In "case1", the Random Forest method provides us a test set accuracy of 84.8%. For this method, I used the "RandomForestClassifier" function from sklearn.ensemble package. And the best parameter I found is 'criterion' = 'Gini', 'max depth'=10, 'n estimators'= 50. In this case, the maximum depth is relatively high, which means the model we created was very complicated. Similar to what I did in the Decision Tree method, a confusion matrix was created for the random forest. The precision is 0.46 and 0.85 for

case	Accuracy	Hyperparameter	precision	recall	F1
case 1	84.80%	max_depth = 10 n_estimator = 50	pos = 0.85 neg = 0.46	pos = 1.00 neg = 0.00	0.46
case 2	63,69%	max_depth = 10 n_estimator = 100	pos = 0.54 neg = 0.64	pos = 0.08 neg = 0.96	0.45
case 3	32.72%	max_depth = 10 n_estimator = 150	1 = 0.35, 2 = 0.26 3 = 0.28, 4 = 0.33	1 = 0.71, 2 =0.11 3 = 0.22, 4 = 0.17	0.27
case 4	47.02%	max_depth = 10 n_estimator = 100	1 = 0.48, 2 = 0.00 3 = 0.39, 4 = 0.00	1 = 0.90, 2 =0.00 3 = 0.13, 4 = 0.00	0.21

Table 4: Test results for Random Forest

positive and negative tweets respectively and the recalls are 0.00 and 1.00. For “case 2”, the test accuracy is 63.69% with best parameter of ‘max depth =10’ and ’n estimators’= 100. In “case 3” and “case 4”, the text accuracies are “32.72%” and “47.02%”. The “max depth” for both of these cases were 10, but the number of estimators for “case 3” is 150 while the number of estimators for “case 4” is 100. The corresponding precisions and recalls are shown in the below table 4.

case	Accuracy
case 1	82.04%
case 2	63.45%
case 3	29.22%
case 4	45.88%

Table 5: Test results for Naive Bayesian

4.6.4 Naive Bayesian

I import “GaussianNB” function from package ‘sklearn.naive Bayes’ to run the model. This function applies Gaussian Naive Bayes method for classification. The test accuracy for “case1” is 82.04%.For “case 2”, “case3”, and “case 4”, the test accuracies are “63.45%”, 29.22%, and 45.88% respectively. Compared to other models, it seems that

Naive Bayesian is the model with the lowest test accuracy in the same case. Table 5 lists the tests accuracy for each case.

5 Discussion

I found that the test accuracies vary largely based on the classifications of the dataset. The case with the highest average test result is “case 1”, in which the data was classified by whether the student studied independently. The case with the lowest average test results is “case 3”, in which the data was classified based on the quantile of the student study time. Here below are the discussion for each case.

Case 1

The reason for the highest accuracy among all three cases might be the highly unbalanced distribution of “0” and “1” in the dataset, 15%, and 85% respectively. According to the confusion matrix above, it seems that the models are predicting most of the results as “1”. Since “1” covers the majority of the dataset, the test accuracy for these models is relatively high.

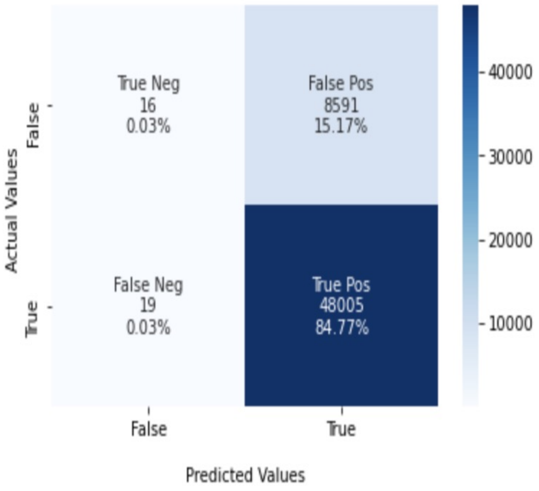

Figure 3: confusion matrix for case 1

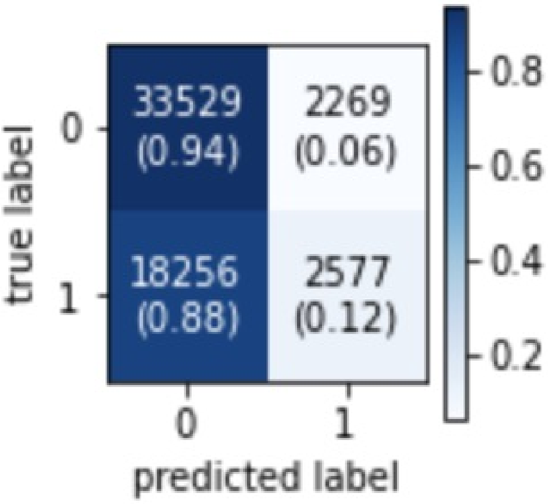

Figure 4: comfusion matrix for case 2

Case 2

For “Case 2”, the dataset is much balance distributed with 119212 0 s and 69556 1s.

However, according to the confusion matrix of the Decision Tree and Random Forest, it seems the models are classifying 90% of the results into 0. This means that, for the 88% of time in the test dataset, even though the individual of the student is above 7 hours per week, the model would still classify the student into the group that contains students who study less than 7 hours per week.

Case 3

"Case 3" used quantile as the standard to label these data, which made sure that every label has enough and balanced samples for the model to make a prediction. Even though the overall test accuracy for the decision tree model is relatively low compared to other case results, the corresponding confusion matrix presents an interesting point. The model is classifying half of the data from each label into "0", which is studying less than 2.5 hours per week. But it seems that the models might had a hard time distinguishing between the higher study time level.

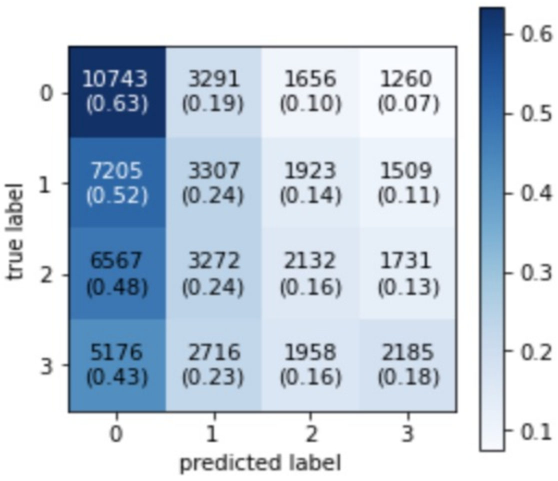

Figure 5: confusion matrix for case 3

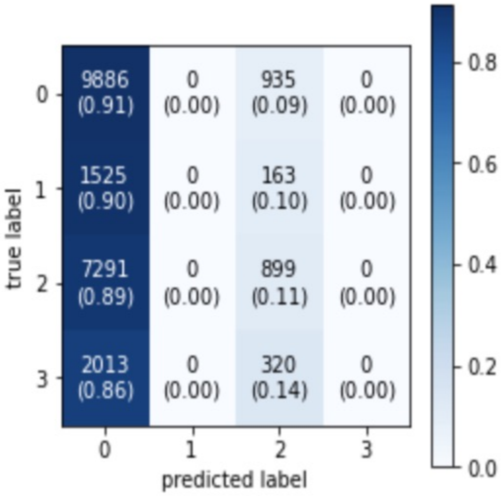

Figure 6: comfusion matrix for case 4

Case 4

The result of exploration for study time above the average is interesting. According to confusion matrix, both of the Decision Tree model and Random Forest model did not classify any data into level 2 and level 4, which are individual study time between

10 and 12 and the individual study time between 20 and 28 respectively. 90% of the predictions are gathered in level 1, which is the individual study time between 10 and 12. This indicates that the models detect no differences between these levels and the classify them into the lowest level.

Summary

One noticeable thing is that decision trees and random forests are producing close accuracies, but the running time for decision trees is much shorter than the running time for the random forests. In this case, to those scholars in the similar research who aim to pursue the efficiency and accuracy of the models, I recommend using the decision tree method instead of the random forest method, as the decision tree model is more effective than random forest method in these cases. Also, Naive Bayesian is the model with the lowest accuracies, indicating the model might not be suitable in predicting such a complicated dataset.

These results indicate that the models are good at predicting binary outcomes like whether the student studied or not, or whether the individual study time for the student is above the average. However, the model may have a hard time making further predictions on the level of the time student spends independently. This can be seen in confusion matrix of "case 3" that the prediction accuracy for each level were very close to each other. The pattern is more obvious in "case 4". In this case, we could conclude that the features chosen in this research cannot make too many contributions on identifying the level of individual study time for each student. We could interpret this in the following way: when the study time is not 0, other features or parental factors may not clearly impact the level of individual study time. In order words, as long as the student has the intent to study, the family background and related factors may not be direct reasons to influence the hours that students spend on studying.

Another interesting thing in this project is that no matter how I separate the dataset, the feature importance stays the same. Among these important features, household income appears to be the most significant one. This matches the results of many previous studies that indicate the income of a family is a great factor that could influence education inequity. For example, in William's paper [13], he mentioned that the reason for increasing drop out rate among students from poor families is that the youths have to do extra work to support the family during the Ebola pandemic period, and thus have to drop out from school. This could also be the reason that household income is the feature that influences the outcome the most. In addition, parents' mental health also plays an important role in student's individual study time. According to Brooks et al. [5], the parents who are under quarantine are more likely to get a mental health disorder than the parents who are not. Moreover, the mental health disorder of parents is negatively related to the mental health condition of the children in the household [9]. This indicates that parents' mental health disorder(s) may negatively influence students' motivations for independent study.

6 Conclusion

My goal for this project was to explore factors that could possibly influence the education inequity during the COVID-19 pandemic period. In particular, I examined study time of students as a measurement of potential education inequity. Even though some models are providing high test accuracies, it is worth further splitting the data by different standard to support deeper exploration. In addition, as shown in "case 3" and "case 4", the models had a hard time identifying between the study time with higher level, but I found that the feature importance stay the same no matter how I classify the data. Among those, household income seems to have the greatest influence on student study time, which children in richer households studying for a longer time independently.

There are still lots of improvements that can be made in future research. If possible, we can further narrow the target family to those who have kids taking grade 6-12 education because these children are more likely to have the ability and self-discipline to study alone. Another thing we could do in the future is to follow up with the academic performance of the student to have a longer-term measurement of education inequity in achievement. Also, even though the feature importance part shows that most of selected important features are family factors, we could still explore more on factors that regarding school characteristics such as learning forms that the school provide; the diversification of the school; as well as financial condition of the school.

7 Reflection

In this independent study, my skills in analytic improved a lot. I learned how to search for possible data sets online and how to extract the information that I want from these datasets. I am more comfortable with dealing with large datasets. Also, by implementing and interpreting both classification and linear machine learning models in this project, I have a better understanding of these models. In addition, since I used both R and Python in this study, my skills for processing data and building models with these two programming languages improved significantly.

My skills for writing scientific research were also improved during this project. I learned a lot about how to interpret different models and how to explain the differences between these models. In addition, my ability to connect the result of my research to the results of other research were also improved. However, I would explore more features regarding school factors or psychological factors and also learn and apply the more advanced algorithms and models in the research if I have more time.

Finally, I would like to express my appreciation to professor Ran Liu for being the instructor of my research. I am grateful for her time in guiding me through the whole research.

References

[1] Household pulse survey, https://www.census.gov/data/experimental-data-products/household-pulse-survey.html.

[2] Policy data, https://statepolicies.com/data/.

[3] E. R. Adrian and H. Michael. Maximally maintained inequality: Expansion, reform, and opportunity in irish education. Jan 1993.

[4] M. Caroline and N. Abyshey. The digital divide and higher education challenge with emergency online learning:analysis of tweets in the wake of the covid-19 lockdown. *Turkish Online Journal of Distance Education*, 22(4), 2021.

[5] B. et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. February 2020.

[6] F. et al. U.s. children “learning online” during covid-19 without the internet or a computer: Visualizing the gradient by race/ethnicity and parental educational attainment. *Socius*, 7, 2021.

[7] N. et al. Academic performance of college students: Influence of time spent studying and working. *Journal of Education for Business*, 2006.

[8] A. Gamoran. American schooling and educational inequality: A forecast for the 21st century. *Sociology of Education*, 74, 2001.

[9] S. Ginny and M. Miriam Silman. Posttraumatic stress disorder in parents and youth after health-related disasters. *Disaster Medicine and Public Health Preparedness*, 7(1), 2013.

[10] F. Julia. Troubled schools in troubled times: How covid-19 affects educational inequalities and what measures can be taken. *European Educational Research Journal*, 20(5), 2021.

[11] S. R. Lucas. Effectively maintained inequality: Education transitions, track mobility, and social background effects. May 2001.

[12] S. Raschka. *Python Machine Learning*. Packt, 20015.

[13] C. S. William. Consequences of school closure on access to education: Lessons from the 2013–2016 ebola pandemic. *Int Rev Educ*, April 2021.

A Appendix

The description for the data are based on the data dictionary from the Household Pulse Survey on United States Census Bureau [1].

- Week: Number of week that the observation was included
- UI_quarantined: Whether there is Expanded eligibility of unemployment insurance to anyone who is quarantined and/or taking care of someone who is quarantined. (0 = "No", 1 = "Yes")
- UI_highrisk: Whether there is Expanded eligibility for unemployment insurance to high-risk individuals in preventative quarantine. (0 = "No", 1 = "Yes")
- States: The state of each respondent
- state_metropolitan: Whether the place the respondent in is a state or metropolitan
- Marital_Status: the marital status of each respondent
- Food_Sufficency: whether there is enough food sources for each household (0 = "No", 1 = "Yes")
- Health_Status: Whether the respondent thinks he or she is in a good health status (1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor)
- Private_Health_insurance_status: Whether the respondent has a private health insurance (0 = "No", 1 = "Yes")
- Public_Health_insurance_status: Whether the respondent has a public health insurance (0 = "No", 1 = "Yes")
- Worry: the frequency of worry that the respondent experiences each week (1 = not at all, 2=several days, 3=more than half the days, 4 = nearly every day)
- Depression: the frequency of depression that the respondent experiencing each week (1 = not at all, 2=several days, 3=more than half the days, 4 = nearly every day)
- no_interest: the frequency of having little interest in things that the respondent experiencing each week (1 = not at all, 2=several days, 3=more than half the days, 4 = nearly every day)
- Incomeloss: Whether there are people in respondent's household experiencing job loss (0 = "No", 1 = "Yes")
- Householdincome: the range of household income before tax in 2019 (1 = less than 25000, 2 = 25000 – 34999, 3 = 35000 – 49999, 4 = 50000 – 74999, 5 = 75000 – 99999, 6 = 100000 – 149999, 7 = 150000 – 199999, 8 = more than 200000)
- teach_cancelled: whether class canceled during the pandemic period (0 = "No", 1 = "Yes")

- teach_distance_online: whether class changed into distance learning using online materials (0 = ”No”, 1 = ”Yes”)

- teach_distance_paper: whether class changed into distance learning using paper materials sent home to children(0 = ”No”, 1 = ”Yes”)

- teach_others: Whether class changed into other forms(0 = ”No”, 1 = ”Yes”)

- teach_no.change: Whether class has no changes because the school did not close(0 = ”No”, 1 = ”Yes”)

- teach_livecontact: children’s virtual contact with teachers each week (1 = none, 2 = 1 day, 3 = 2-3days, 4 = 4 or more days)

- teach_indivisualstudy: children’s hours spent studying on their own each week (numerical)

- teach_studytime: children’s hours spent on studying each week (1= Much less than a school day before the coronavirus pandemic, 2 = A little bit less than a school day before the coronavirus pandemic, 3 = As much as a school day before the coronavirus pandemic, 4= A little bit more than a school day before the coronavirus pandemic, 5= Much more than a school day before the coronavirus pandemic)

- Health_insurance: whether the respond has insurance or not. (0 = ”No”, 1 = ”Yes”)